

Machine Learning 1

Lecture 2.6 - Bayesian Prediction

Erik Bekkers

(Bishop 1.2.6)



Slide credits: Rianne van den Berg

Image credit: Kirillm | Getty Images

Bayesian Approach

$$p(t' | x', \underline{w}, \beta)$$

- Dataset $D = (x_1, x_2, \dots, x_N)$ of N independent observations.

$$(t_1, t_2, \dots, t_N)$$

- Frequentist approach: search for one optimal estimate of \mathbf{w}

$$\mathbf{w}_{\text{ML}} = \underset{\underline{w}}{\operatorname{argmax}} p(D | \underline{w})$$

$$\mathbf{w}_{\text{MAP}} = \underset{\underline{w}}{\operatorname{argmax}} p(\underline{w} | D)$$

- Bayesian approach: Given a prior belief over \mathbf{w} , $p(\mathbf{w})$, and our data D , we are interested in the posterior distribution

$$p(\mathbf{w} | D) = \frac{p(D | \underline{w}) p(\underline{w})}{p(D)}$$

- $p(\mathbf{w} | D)$ reflects the plausibility of different \mathbf{w} , given our prior knowledge and how likely our data is generated using \mathbf{w} .

Bayesian Approach

- ▶ Prior distribution: $p(\mathbf{w})$, should represent some prior knowledge/belief of the plausibility of \mathbf{w} .
- ▶ After observing data $D = (x_1, x_2, \dots, x_N)$, posterior distribution

$$p(\mathbf{w} | D) = \frac{p(D | \mathbf{w})p(\mathbf{w})}{p(D)}$$

- ▶ Predictive distribution:

$$\begin{aligned} p(x' | D) &= \int p(x', \mathbf{w} | D) d\mathbf{w} = \int p(x' | \cancel{D}, \underline{w}) p(\underline{w} | D) d\underline{w} \\ &= \int p(x' | \underline{w}) p(\underline{w} | D) d\underline{w} \end{aligned}$$

Conditional independence

► **Note:** even if $p(D | \mathbf{w}) = \prod_{i=1}^N p(x_i | \mathbf{w})$

$$p(D) = \int p(D, \mathbf{w}) d\mathbf{w} = \int \underbrace{p(D | \mathbf{w})}_{\text{blue bracket}} \underbrace{p(\mathbf{w})}_{\text{blue bracket}} d\mathbf{w} \neq \overbrace{\prod_{i=1}^N p(x_i)}^{\text{red bracket}}$$

$$\int p(x_1 | \underline{w}) p(x_2 | \underline{w}) \dots p(x_N | \underline{w}) p(\underline{w}) d\underline{w}$$

$$\neq \int p(x_1 | w_1) p(w_1) dw_1 \int p(x_2 | w_2) p(w_2) dw_2 \dots$$

Curve Fitting: Bayesian Approach

► Dataset $D = \{(x_1, t_1), \dots, (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

► Posterior distribution after observing data:

$$p(\mathbf{w} | \mathbf{x}, \mathbf{t}) = \frac{p(\underline{t} | \underline{x}, \underline{w}) p(\mathbf{w})}{p(\underline{t} | \underline{x})} \quad \text{with} \quad p(\mathbf{t} | \mathbf{x}) = \int p(\mathbf{t} | \mathbf{x}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}$$

► Predictive distribution:

$$\begin{aligned} p(t' | x', \mathbf{x}, \mathbf{t}) &= \int p(t', \mathbf{w} | x', \mathbf{x}, \mathbf{t}) d\mathbf{w} \\ &= \int p(t' | x', \cancel{x}, \cancel{t}, \underline{w}) \cdot p(\underline{w} | \underline{x}, \underline{t}) d\underline{w} \\ &= \int p(t' | x', \underline{w}) \cdot p(\underline{w} | \underline{x}, \underline{t}) d\underline{w} \end{aligned}$$

Curve Fitting: Bayesian Approach

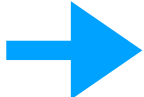
- ▶ Predictive distribution: $p(t'|x', \mathbf{x}, \mathbf{t}) = \int p(t'|x', \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})d\mathbf{w}$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{x})} \quad \text{with} \quad p(\mathbf{t}|\mathbf{x}) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w})d\mathbf{w}$$

Advantages:

- ▶ Inclusion of prior knowledge
- ▶ Represents uncertainty in t' both due to target noise, and uncertainty over \mathbf{w} .

Disadvantages:

- ▶ Posterior is hard to compute analytically  approximate!
- ▶ Prior is often chosen for mathematical convenience, not reflection of prior belief!